## REMARKS/ARGUMENTS

Reconsideration of this application is respectfully requested.

The Examiner is thanked for including "a Response to Arguments" section at pages 20-22. As the Examiner recognizes, Applicants have argued that while Jones does teach a dynamically generated table of contents (TOC) for the static content of a website, Jones does not appear to have relevant teaching with respect to generating a TOC for dynamically generated webpage content, etc. In rebuttal, the Examiner alleges that Jones does teach generation of dynamically generated network pages and relies upon column 3, lines 19-20:

> "The present invention provides a method and apparatus for navigating through electronically stored information using an expandable, hierarchial index or TOC, in a hyper-textual client-server network environment such as the world wide web"....[3: 19-23]

The Examiner also relies upon column 3, lines 38-39 to demonstrate that the alleged dynamically generated network page is in servable form. However, it appears that this portion of Jones is also referring only to the dynamically-generated table of contents:

> "...upon receiving the network request, and based upon the address path and the digital specification, the server dynamically generates a network page specifying display of a hierarchial portion of the TOC entries. This network page is transmitted from the server to the client, for display to the end-user." [3: 34-39]

The Examiner has also criticized independent claim 49 and dependent claim 58 as not being sufficiently specific to support the argument that the Jones reference to scripts, data files, data bases, etc. are actually part of the Jones system itself rather than a part of the network site being analyzed and thus external to the claimed system. The Examiner also indicated the need for more specificity in the claims with respect to "valid parameters".

Accordingly, the claims have now been amended to a form that is, hopefully, more clearly distinguished from Jones. If the Examiner has further criticisms of the claim language vis-à-vis, arguments of distinction, etc., it is requested that the undersigned be telephoned so that prompt resolution of any such issue can be effected.

Unfortunately, the Form PTO/SB/08a returned with the last Office Action was only partially initialed. No initials, lines through, or other notations of any kind were made for the two "Other Documents" listed on this form. Indeed, the International Search Report (ISR) was not only filed with this Information Disclosure Statement (IDS), but was also directly forwarded to the U.S. Patent and Trademark Office by WIPO and the U.S. Patent and Trademark Acknowledgement of same is explicitly stated in the Acceptance Letter mailed 05/16/2005. Nevertheless, a further copy of both the ISR and the printed publication document listed are now attached.

Return of a fully initialed copy of the relevant Form PTO/SB/08a is respectfully requested.

The Examiner's attention is also directed to a further IDS filed March 5, 2008. Return of a fully initialed copy of that Form PTO/SB/08a is also respectfully requested.

The rejection of claims 49-50, 52-58, 68-69, 71, 72, 75, 77, 78, 91, 94, 97 and 99-110 under 35 U.S.C. § 102 as allegedly anticipated by Jones '098 is respectfully traversed.

By way of introduction, Jones describes a CGI script that dynamically generates web pages representing a table of contents, or "TOC", for content of a web site. However, the TOC pages in Jones are generated from a single fixed "structure definition file" that pre-defines the TOC structure and headings, and includes every hyperlink to corresponding content (i.e., web pages) that are included in the TOC. The "structure definition file" in Jones is manually

prepared by a user, who therefore must select the web pages that are to be included in the TOC, and also choose the desired structure and headings that are to be used.

Given this pre-determined structure and format, the CGI script then dynamically generates web pages representing the TOC pages as a user navigates through the TOC. To access site content, a user selects one of the hyperlinks to site content, which are provided as leaf nodes to the TOC hierarchical structure, such as the hyperlink 140 shown in Figure 1E.

There is no teaching or suggestion in Jones that the site content linked to by the TOC leaf nodes could include dynamically generated content. In the outstanding Office Action, the Examiner continues to assert that the TOC pages themselves are part of the "content" of the network site, although it is believed that a person skilled in the art would not construe the term in this way.

Additionally, in Jones, the requests issued do not appear to include parameters, but rather a URL without parameters, consisting only of a protocol identifier and a domain name. On the other hand, perhaps one could argue that the path components in the URL following the CGI script name are effectively parameters, since the script uses those components to determine its behavior. For example, in Jones, the link target "niftnavi.cgi/ST/STCH" shows at column 9 line 49 is used to determine array members "ST" and "STCH", as described at column 7 lines 63 to 67.

In view of this ambiguity, Applicant has now presented amended (and new) claims to more clearly distinguish from Jones. For example, claim 49 has been amended to now require the generation of "data representing nodes of a hierarchy of linked nodes for indexing content of said network site, wherein leaf nodes of said hierarchy include alternative links for use in accessing said dynamically generated content."

In Jones, leaf nodes are standard hyperlinks to static content. There is nothing in Jones to teach or suggest that leaf nodes could be alternative links for use in accessing corresponding dynamically generated content.

These features are also included in new process claim 118.

It will also be noted that the requirement of indexing has been removed from the independent claims. The invention is believed to have broader application. However, the recited portion of claim 49 includes the words "a hierarchy of linked nodes *for use in indexing content of said network site.*" This recitation does not require actual indexing, but only that the hierarchy of linked nodes is *capable of* being used for indexing site content.

Claim 49 also now requires processing one of the alternative links to determine a corresponding link and one or more corresponding parameters that, in combination with the corresponding link, determine corresponding dynamic content of the network site. This feature is also recited in new independent claim 160.

Given such fundamental deficiencies of Jones with respect to independent claim 49, it is not believed necessary at this time to explain further deficiencies of Jones with respect to additional features of independent claim 49 or of the rejected dependent claims.

The rejection of claims 79-86, 92 and 95 under 35 U.S.C. § 102 as allegedly anticipated by Jones is also respectfully traversed. However, since these claims have been canceled without prejudice or disclaimer, this ground of rejected has been mooted and it is not believed necessary therefore to discuss it in further detail at this time.

The rejection of claims 89, 90, 93 and 96 under 35 U.S.C. § 102 as allegedly anticipated by Jones is also respectfully traversed. Once again, these claims have been canceled without

prejudice or disclaimer, thus mooting this ground of rejection and making further discussion at this time unnecessary.

The rejection of claim 51 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Steele '737 is respectfully traversed.

Fundamental deficiencies of Jones with respect to amended parent claim 49 have already been noted above. Steele does not supply those deficiencies. Accordingly, it is not believed necessary at this time to discuss the additional deficiencies of this allegedly "obvious" combination of references with respect to claim 51 or its parent claims.

The rejection of claims 59 and 60 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Cochran '934 is also respectfully traversed.

Fundamental deficiencies of Jones have already been noted above with respect to parent claim 49. Cochran does not supply those deficiencies. Accordingly, it is not believed necessary at this time to discuss additional deficiencies of this allegedly "obvious" combination of references with respect to the additional features of claims 59 and 60.

The rejection of claim 61 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in further view of Cochran is also respectfully traversed -- for reasons already discussed above with respect to claims 59 and 60.

The rejection of claim 62 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Steele '737 is also respectfully traversed.

Once again, fundamental deficiencies of Jones have already been noted above with respect to parent claim 49. Since Steele does not supply those deficiencies, it is not believed necessary at this time to discuss additional deficiencies of this allegedly "obvious" combination of references with respect to claim 62.

The rejection of claims 63 and 64 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Conner '152 is also respectfully traversed.

As before, fundamental deficiencies of Jones have already been noted above with respect to parent claim 49. Conner does not supply those deficiencies. Accordingly, it is not believed necessary at this time to discuss additional deficiencies of this allegedly "obvious" combination of references with respect to dependent claims 63 and 64.

The rejection of claim 65 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Salemo WO '463 is also respectfully traversed.

Once again, fundamental deficiencies of Jones have already been noted above with respect to parent claim 49. Salemo does not supply those deficiencies either. Accordingly, it is not believed necessary at this time to discuss additional deficiencies of this allegedly "obvious" combination of references with respect to the additional features of dependent claim 65.

The rejection of claim 66 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in further view of McCormack '680 is also respectfully traversed.

As noted above, there are fundamental deficiencies of Jones with respect to parent claim 49. McCormack does not supply those deficiencies. Accordingly, it is not necessary at this time to explain further deficiencies of this allegedly "obvious" combination of references with respect to the additional recitations of dependent claim 66 -- which must be considered "as a whole" under 35 U.S.C. § 103.

The rejection of claim 67 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Steele '737 is also respectfully traversed -- for reasons already noted above with respect to allegations concerning other dependent claims where neither Jones, nor Steele, nor any combination thereof teaches or suggests the recitations of parent claim 49.

Accordingly, it is not necessary to discuss the additional deficiencies of this allegedly "obvious" combination of references with respect to the additional recitations of claim 67 which must be considered "as a whole" with its parent claim under the provisions of 35 U.S.C. § 103.

The rejection of claim 70 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Steele '737 is also respectfully traversed -- for reasons already noted above. With respect to other claims dependent from patentably distinguished parent claim 49.

The rejection of claims 73 and 74 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Steele '737 is similarly traversed -- for reasons already noted above. With respect to other claims dependent from parent claim 49.

The rejection of claim 76 under 35 U.S.C. § 103 as allegedly being made "obvious" based on Jones in view of Steele '737 -- and the similar rejection of claims 87 and 88 based on Jones in view of Steele are all respectfully traversed -- for reasons already noted above with respect to other claims dependent from patentably distinguished parent claim 49.

It is also respectfully noted that the alleged "obviousness" of combining various selected bits and pieces from numerous other references with respect to Jones itself demonstrates non-obviousness under 35 U.S.C. § 103. But for the Applicants' teaching, where would any person of only ordinary skill in the art at the relevant time have found any motivation, suggestion, or teaching to make the numerous and different selective piecemeal combinations now being alleged as "obvious"? Nor would one having only ordinary skill in the art find any "common sense" or other reason for making such combinations in the absence of the Applicants' own teaching.

The Examiner's attention is drawn to new claims 111-160. New claims 111-117 depend directly or indirectly from claim 49. New claim 118 is based on claim 49 as well, but without

the "proxy" portion now added as the last two elements of claim 49. New dependent claims 119-152 track claims dependent from claim 49, except that these claims now depend from claim 118.

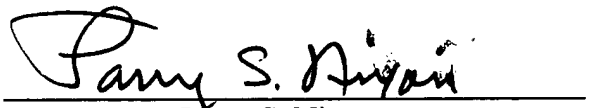
New independent claim 153 includes "proxy" recitations similar to those now found in claim 49 with a new "indexible" limitation. That is also believed to be patentably distinct. Claims 154-159 depend directly or indirectly from claim 153.

New claim 160 is a computer system claim which requires, *inter alia,* a network site analysis component which is capable of handling dynamically generated content (i.e., generated in response to receipt of at least some requests including parameters that determine the dynamically generated content and then to analyze the accessed files to identify at least one valid parameter, etc.). Claim 160 also requires a link generator to generate, based on such analysis, data representing nodes of a hierarchy of link nodes for use and indexing content of the network site wherein leaf nodes of the hierarchy include alternative links for use in accessing the dynamically generated site content. These and other recitations in independent claim 160 clearly distinguish from any teaching of the cited references. New claim 161 depends from claim 49 and requires a hierarchy to be determined by servable content of the network cite and at least one valid parameter value, etc., used by executable code and/or scripts at the network site to determine the dynamically generated content.

Accordingly, all now presented claims are now believed to be patentably distinct from any teaching or suggestion of the cited references. The formal notice to that effect is respectfully solicited.

Respectfully submitted,

**NIXON & VANDERHYE P.C.**

By: 

Larry S. Nixon
Reg. No. 25,640

LSN:lmj
901 North Glebe Road, 11th Floor
Arlington, VA 22203-1808
Telephone: (703) 816-4000
Facsimile: (703) 816-4100

INTERNATIONAL SEARCH REPORT

International application No.
PCT/AU03/00904

**A. CLASSIFICATION OF SUBJECT MATTER**

Int. Cl. [7]: G06F 17/30

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
USPTO, IEEE, esp@cenet: internet, web, 'search engine', dynamic, 'invisible web', TOC, spider, agent

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | King-Ip Lin et al., *Automatic Information Discovery from the "Invisible Web"*, International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, April 08-10 2002<br>See whole document | 1, 13, 18, 28, 30, 31, 33, 37 |
| X<br>Y | WO-01/46856-A1 (YOURAMIGO PTY LTD) 28 June 2001 (28-06-01)<br>See whole document | 1-37<br>17-25 |
| Y | US-6199098-B1 (JONES et al.) 6 March 2001 (06-03-01)<br>See whole document | 17-25 |

[X] Further documents are listed in the continuation of Box C [X] See patent family annex

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 3 September 2003 | 9 SEP 2003 |
| Name and mailing address of the ISA/AU<br>AUSTRALIAN PATENT OFFICE<br>PO BOX 200, WODEN ACT 2606, AUSTRALIA<br>E-mail address: pct@ipaustralia.gov.au<br>Facsimile No. (02) 6285 3929 | Authorized officer<br>MICHAEL HARDY<br>Telephone No : (02) 6283 2547 |

INTERNATIONAL SEARCH REPORT

International application No.
PCT/AU03/00904

C (Continuation). **DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | WO-02/07013-A2 (BIAP SYSTEMS, INC.) 24 January 2002 (24-01-02)<br>See abstract and claims | |

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
**PCT/AU03/00904**

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

| Patent Document Cited in Search Report | | Patent Family Member | | | | | |
|---|---|---|---|---|---|---|---|
| WO | 200146856 | AU | 200123285 | CA | 2394820 | EP | 1250663 |
| US | 6199098 | NONE | | | | | |
| WO | 200207013 | AU | 200173591 | CA | 2416182 | EP | 1307832 |
| | | US | 2002026462 | US | 2003004961 | WO | 2003007189 |

END OF ANNEX

# Automatic Information Discovery from the "Invisible Web"

King-Ip Lin, Hui Chen
Division of Computer Science
The University of Memphis,
Memphis, TN 38152, USA
linki@msci.memphis.edu, huichen@memphis.edu

**Abstract**

A large amount of on-line information resides on the invisible web – web pages generated dynamically from databases and other data sources hidden from the user. They are not indexed by a static URL but is generated when queries are asked via a search interface (we denote them as specialized search engines). In this paper we propose a system that is capable of automatically making use of these specialized engines to find information on the invisible web. We describe our overall architecture and process: from obtaining the search engines to picking the right engines to query. Experiments show that we can find information that is not found by the traditional search engines.
***Keywords:*** *Search engines, invisible web*

## 1. Introduction

The World Wide Web can be viewed as a huge repository of diverse information. For instance, Google [2] estimated that it has access to 1.3 billion web pages at August 2001 [11]. One can think of the Web as a huge encyclopedia with coverage on every subject that is "out-there".

With a large repository of the data, the challenge is to provide a mechanism for any user to effectively retrieve information that he/she is looking for. For this purpose, a multitude of *search engines* have sprung up to retrieve information over the Web. Examples include *Google, Yahoo, Infoseek,* and *Altavista*. In general, these engines crawl the Web by following URLs that are located on the Web pages that they encountered. The engines store and index all the Web pages encountered into their local databases –using URL and other keyword information. When a request – usually in the form of keywords – arrives, the local databases are searched and the appropriate web pages are returned.

Recently, however, there has been a focus on a part of the Web that is untouched by the traditional search engines. Known as the "invisible web" or "deep web", it represents information on the web that are not indexed by the regular search engines. A lot of information that appears on the web is stored in specialized databases. These information do not appear in static web pages, but are accessed through specific search interfaces (using CGI and HTML forms or Javascript, for instance) – which we called specialized search engines. These databases cannot be indexed and searched by traditional engines, as they do not have a static URL for the traditional search engine to index. However, they are a goldmine of information, as many databases contain detailed and specific information that are not present in other parts of the web. For instance, a study in 2000 done by brightplanet.com [6] suggested that the invisible web contains about 400-550 times the information of the traditional, indexable World Wide Web – which adds up to about 550 billion documents and 7,500 terabytes of information. Ability to access the invisible web will be a tremendous boost for information retrieval over the Web. Another advantage of the invisible web is that each database contains data from a specific domain (from car prices to court cases). When the user's need fit the domain of a certain specialized engine, it is likely that highly relevant information can be obtained.

There has been work on building tools to allow users to access the invisible web. Many of them are directories of the specialized search engines for the invisible web. They include invisibleweb.com [5] (from Intelliseek) and direct search [1] (from George Washington University). They have been building tools to collect the web pages containing those specialized search engines, and provide the interface to allow users to access those engines. However, there are some limitations on the service that these systems provide:

- Manual search: Most systems allow users to browse their collection of engines. Some allow users to search for the appropriate engine (based on keywords from the web page where the search engine interface resides). However, currently these systems do not automatically ask the queries on the specialized search engine itself. One reason is that it is not clear what is the right query phrase

Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'02)
0-7695-1506-1/02 $17.00 © 2002 IEEE
IEEE COMPUTER SOCIETY

to be sent to the specialized search engines. For instance, some search engines may be looking for names, while others are looking for social security numbers. Also, different search engines have different interfaces, thus there is no one standard way of sending the right query. Thus all searches there has to be manual, and the systems above provide little help in that regard.

- Query specification: Currently, all search engines let users search by keywords. Users supply a set of keyword(s) and the system matches the keywords to the web pages. This method of searching depends on the users supplying the right keywords. Many situations, from insufficient domain knowledge to words that have multiple meanings, can lead to the poor choice of keywords and the search engine failing to find the right results.

In this paper, we describe our approach in searching the invisible web. We propose a system that maintains information about the specialized search engines in the invisible web. When a query arrives, the system not only finds the most appropriate specialized engines, but also redirects the query automatically so that the user can directly receive the appropriate query results. Our system has the following characteristics:

- **Database of specialized search engines.** Our system maintains a database of specialized search engines, storing information such as URL, domain, and search fields. The information allows the system to pick the right search engine to access, as well as automatically construct the query to be sent.
- **Automatic search engine selection.** The system utilizes the search engine database to automatically decide which search engine(s) to use and route the query appropriately.
- **Data mining for better query specification and search.** We apply data mining techniques to discover information related to the search keywords so as to facilitate both the search engine selection and query specification process. This enables more relevant results to be returned.

We have implemented the major components of the system and have tested them. We believe that our techniques can provide a better search tool for the multitude of the information over the Internet.

The rest of the paper is organized as follows. Section 2 outlines our system – its architecture, as well as the search algorithms. We present experimental results on section 3. Section 4 compares our work with other related work. We conclude our paper with potential future work in section 5.

## 2. System overview

As stated, the goal of our system is to enable access the information of the invisible web, via automatically utilizing the specialized search engines. Our system is divided into four components:

**The crawler**: This is used to populate the database with specialized search engines, as well as extracting the information that need to be stored with each one of them.

**The search engine database**: This is where the information about the individual specialized search engine is stored.

**The query pre-processor**: This component receives user input and finds phrases that associate with the query keywords. These phrases form the basis of the search engine selection and ranking system.

**The search engine selector**: This component receives the information from the pre-processor as well as the search engine database, and selects the specialized search engines to route the query to, as well as set up the appropriate query to each of them.

Figure 1 outlines the overall architecture and process of our system. It is divided into five steps: populating the search engine database; query pre-processing; search engine selection; actual query execution and result post-processing. The following subsections describe each step in more detail.

### Populate the search engine database

Before one can ask a query, the system needs to populate the database with information of the specialized search engines. This is done by using a crawler to crawl the World Wide Web. However, instead of storing every Web page it encounters, it only extracts the portion that corresponds to a search engine. This is done by locating the "form" tags of the web pages – as the "form" tags denote that user input is expected. In some cases, a web page can have many "form" tags. We treat the different form tags as separate specialized search engines.

For each search engine, we have three fields associated with it:

*Engine description:* This field contains a description of the function of this engine. Certain meta-tags from the engine's web page, like "description", "copyright", "author" provide description about what the engine is about.

Another source of information for the description are "back-links" – other web pages that points to the Web page where the search engine resides. The description (especially the text associated with the URL links) can provide further clues about what this search engine's function is. We clean the content of these


IEEE COMPUTER SOCIETY

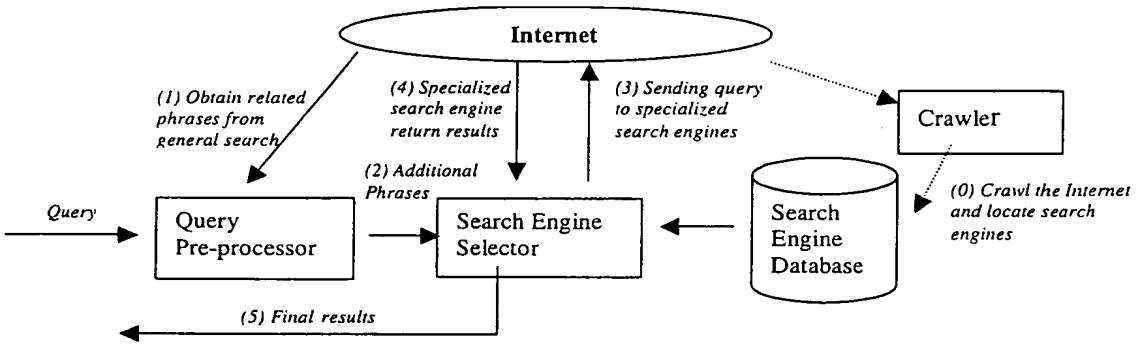



**Figure 1: Overall system architecture**

pages by removing stop words (such as articles and auxiliary verbs).

*Web-page keywords:* we capture keywords that, while not formally in one of the description field, may provide information about the search engine. These include value of title tag and value of attribute "title", "keywords", "keyword", "page title", "page abstract", and "special" in "meta" tags of the web page.

*Engine-wise keywords:* we also store the words and phrases around the search form tag itself. For instance, phrases such as "enter your first name", "enter the e-mail address" can provide further information about the search engine itself, such as the type of search words that is expected.

The main reason we divide the information into three parts is to evaluate which portion is more effective in helping the system to locate the right search engines.

All this information can be obtained automatically during crawling. For each search engines, apart from the three fields mentioned, the system stores its URL and other information necessary to automatically construct the correct query string to send to the search engine.

## Query pre-processing

The goal of this step is to supplement the keywords supplied with the user with other information that can be helpful to determine which engine to search and what keyword to send. To achieve this, we need some kind of knowledge base about keywords and what other words and phrases are associated to it. While standard sources such as WordNet [4] may be useful, the web itself may supply us the best source of information. Thus we take the following approach:

Send the keywords to some general search engines (like yahoo or Google) for a query and return the top results. Based on the results, find words and phrases that appear often with the search keywords.

For the second step, we apply a technique very similar to the one proposed by Mohonen et al. [7], using episode rules discovery techniques [8] to find phrases that associate with the keywords, or phrases that appear often in the result set obtained from the general search engines.

Moreover, certain things, like e-mail addresses, can prove useful for certain search engines (for instance, some search engines can locate information about e-mail address). Our pre-processing step also extracts this information and sends it to the next step.

## Engine selection

The heart of our system is the selection of the search engines to route our queries to. At the end of the previous step, the system has, in addition to the user query, an extra set of keywords and phrases that is used here to determine which search engine to use.

Currently the system applies a keyword/phrase matching approach: each keyword/phrase generated from the pre-processing step is matched with the three fields of each search engine in the database. We rank the number of matches for each search engine and return those with the most matches.

Since each engine has three fields containing different information in the database, we can assign different weight to each of these fields, so that some information carries more importance during the matching process.

## Query execution and result post-processing

After the search engines are selected, the system automatically sends the query to all the search engines and awaits the results to return. Based on the information stored in the database, the system can automatically


IEEE COMPUTER SOCIETY

generate the query string and send the appropriate queries to the web sites. Currently our system is limited to handling queries that use HTML tags. In future we plan to add capabilities to handle other search engines like Javascript based search engines.

One interesting question is what should be the exact query that should be sent to the individual engines. Currently we are simply using the query phrase given from the user. However, in the pre-processing step we obtained other keywords/phrases that is related to the query phrase. We plan to incorporate those words into our query in our next version.

After all the results are received, they are combined and sent to the user interface. Currently we just simply return the links to the resulting web pages. However, in the future we plan to apply other techniques such as clustering to organize the results in a fine manner.

## 3. Experimental results

In this section we present some preliminary experiment results. In the experiments we want to find the best matching process, as well as determine our system's performance as compared to other general search engines.

### Experimental setup

We crawl the web to populate our database with search engines. To aid in crawling, we access the Google web directory to locate feasible starting point for crawler. Our database has 266 specialized search engines from 144 web sites, covering 15 categories such as Arts, Business, Society, Sports and others.

All the experiments described in this section are carried out on a 333 MHz PC with 128 Mbyte memory.

#### Finding the right search engines

The first thing we check is that whether our ranking method provide us with the appropriate search engines. In order to do that, we pick out certain keywords from each category of the search engines that we store. We expect our method will locate the search engines for the corresponding topics accordingly. We also examine some search engines manually and determine what kind of queries is appropriate for them. We designated 35 such queries and submit them to our system and check whether the expected search engine is returned. Phrases that we used for search include: "3d architecture", "Laser vision surgery", "Sports radio stations" and so on.

In order to compare the usefulness of the three pieces of information associated with the search engines (engine description, web-page keywords and engine-wise keywords), we match the search words each of them individually, and add the score to obtain an overall ranking for the search engine. We apply different weights in order to determine which piece of information is more important. Figure 2 highlights the results.

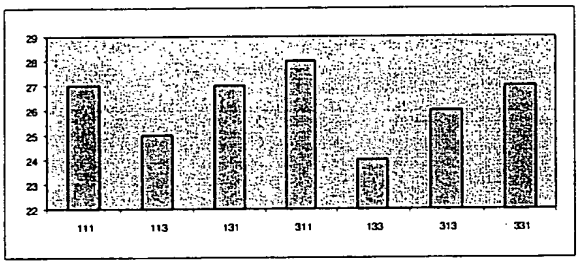



**Figure 2: Performance of matching with various weights**

In Figure 2, each entry corresponds to the weight that is assigned to the engine description, web-page keyword and engine-wise keyword respectively. For each entry, we tabulate the number of cases that the appropriate search engine appear in the top 10 of the results from our algorithm. The number is shown on the y-axis. From the figure, we can see that out of the 35 queries we sent, the expected search engine turned up at the top 10 in more than 70% of the times in most cases. We can also see that there is no significant difference in results when the weights are adjusted, although the performance tends to be better in cases where the description file has a larger weight.

#### Usefulness of the invisible web

Our main goal is to use the specialized search engine to discover information that cannot be obtained by the typical search engines. In order to evaluate the effectiveness of the system for this purpose, we compare the web pages obtained from our search engine to the ones accessed via Google, a well-established search engine over the web. We randomly selected 24 queries phrases (different from the last experiment) and search via our method and Google. For each query, we obtained the top 3 links from our system and compare with the corresponding search results from Google to see if the web page obtained from us is obtainable from Google. Figure 3 tabulate the results:

From figure 3, we can see that 68% of the results contain some related search engines (notice that here we do not pick keywords to ensure search engines can be found; also in many cases there are only 2-3 search engines in our database that relate to search phrases). With a larger database we anticipate a higher percentage of relevant search engines.


IEEE COMPUTER SOCIETY

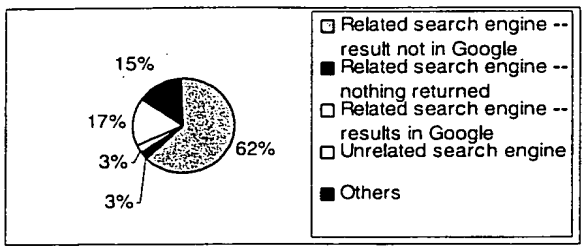



**Figure 3: Comparing search result with Google.com**

In the case where our system returns results, we obtained related and appropriate information from our search. Moreover, virtually every time we are able to access Web pages that are not directly accessible via Google. (The URL is not in Google's result set for this search). Only in 3% of the cases do Google have directly access of the Web page that we returned. While Google's search result do contain relevant information about the search phrase, our system can locate information that is hidden from them.

As an aside, 15% of the results are classified under "others" in figure 3. They are forms that do not correspond to search engines. They are mainly subscription forms ("Enter your e-mail to receive free newsletter!"); thus one challenge for our system is to found out way to filter out such engines during the crawling process.

## 4. Related work

There has been work in building tools to access the invisible web. One approach that has been proposed is the Q-Pilot project [10]. They proposed a similar architecture of storing the search engines and search for the results. However, our approach has some important difference from theirs. We emphasize on automatically locating and collecting information about the search engine during crawling. Also, we distinguish the types of information that are stored for each engine. It seems that the engine-description via the meta-tags provides better results.

As stated in the introduction, there have been some systems implemented that maintain information about the invisible webs. Most of them are directories of specialized search engines with some search capabilities for picking which engines to search (but do not send a query to them). Some publicly available engines include Direct Search [1] and the Invisible Web from Intelliseek. Other systems like deep web query and lexibot [3] are propriety products.

Another project that is related to our work is the HiWE (Hidden Web Exposer) project [9] from Stanford University. They proposed a generic model of a hidden web crawler, as well as a human-assisted technique to extract the information. While their methods are innovative and can be very useful, the large amount of human assistance required can be a drawback. Moreover, they have not explored on how to pick the right search engine to forward the queries.

## 5. Conclusion and future work

In this paper we presented a system that can automatically index and search the invisible web. We show that our system can locate information that are not accessible from general web search engines. Moreover, initial experiments shows that our ranking techniques provide users a good set of specialized search engine to deliver the right information.

We plan to improve our system in many ways: for instance, cleaning of the databases; sending more appropriate queries to the specialized search engines so that the user can have a higher chance to find what he/she wants; and improve on the efficiency of the system.

## Acknowledgements

We would like to thank Raghunandan Upparapalli for his help in implementing query pre-processing.

## Bibliography

1. *Direct Search.*, http://gwis2.circ.gwu.edu/~gprice/direct.htm
2. *Google.com.*, www.google.com
3. *Lexibot.*, www.lexibot.com
4. *WordNet: An Electronic Lexical Database*, ed. C. Fellbaum. 1998: MIT Press.
5. *www.invisibleweb.com.*,
6. Bergman, M.K., *The Deep Web: Surfacing Hidden Value*. 2000, BrightPlanet.com,
7. Mahonen, H., et al., *Applying data mining techniques for descriptive phrase extraction in digital documents.* Proc. Advances in Digital Libraries (ADL98), 1998.
8. Mannila, H., H. Toivonen, and A.I. Verkamo, *Discovery of frequent episodes in event sequences.*



IEEE COMPUTER SOCIETY

Data Mining and Knowledge Discovery, 1997. 1(3): p. 259-289.
9. Raghavan, S. and H. Garcia-Molina. *Crawling the Hidden Web*. in *27th International Conference on Very Large Data Bases*. 2001. Rome, Italy.
10. Sugiura, A. and O. Etzioni, *Query Routing for Web Search Engines: Architecture and Experiments*. 2000 WWW Conference, 2000.
11. Sullivan, D., *Search Engine sizes*. The Search Engine Report, 2001.

Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'02)
0-7695-1506-1/02 $17.00 © 2002 IEEE

IEEE COMPUTER SOCIETY